Last time we discussed:

* models and variability
* types of plots
* five-number summary

# Turning a Statistical Double Play

In the previous class we focused on the use of plots to summarize data, especially plots that have numerical data for both axes. Another tool of descriptive statistics is the *frequency table*, which is especially useful for categorical data. Frequency tables are more than just lists of numbers—they provide information that allows the data to be generalized and easily compared to other data or to models. There are two types: one-way and two-way frequency tables. Although both are useful, the two-way frequency table, once mastered, becomes a powerful method for testing models.

**One-Way Frequency Table**

A one-way frequency table lists data that can be **summed in one direction**.

It can also be used to list the *relative frequency*, *cumulative frequency*, and the *percentage* for each category.

**Activity 1:** Favorite colors of 1st graders (redux)

Based on previous data, how many students in a class of 30 would have each color as their favorite?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Color** | **Frequency** | **Relative Frequency** | **Cumulative Frequency** | **Percentage** (%) | **Prediction** (class of 30) |
| Blue | 6 | 0.30 | 0.30 | 30 |  |
| Red | 4 | 0.20 | 0.50 | 20 |  |
| Green | 3 |  |  |  |  |
| Pink | 2 | 0.10 |  | 10 |  |
| Purple | 2 |  |  |  |  |
| Yellow | 2 |  |  |  |  |
| Orange | 1 |  |  |  |  |
| **Total** | 20 | 1.0 |  | 100 | 30 |

**Frequency**—number of occurrences of each category. Column sums to the total number of occurrences.

**Relative Frequency**—frequency divided by total. Column sums to 1.0.

**Cumulative Frequency**—sum of all previous and current relative frequencies. Column is not summed.

**Percentage**—relative frequency multiplied by 100%.

**Question:** What model is used to make the prediction for the class of 30 students?

**Activity 2:** Mixed-Up Baseball Team

On a baseball team most players play both offense (batting) and defense (playing in the field). One interesting idea to investigate is how offensive production varies by field position. The positions are pitcher, catcher, 1st base, 2nd base, shortstop, 3rd base, left field, center field, right field, designated hitter (AL), and pinch hitter (no field position). For this example, offensive production will be indicated by a simple statistic called *runs tallied* (RT), which is the average of the runs scored and runs batted in (RBI) for each player, i.e. RT = (Runs + RBI)/2. The values given are the average runs tallied in 2013 for each position prior to the All-Star break extrapolated over a full 162-game season.

*Model data* *MLB data*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Position** | **Model** | **Revision?** |  | **Position** | **Runs Tallied** |  |
| Pitcher |  |  |  | A | 74 |  |
| Catcher |  |  |  | B | 68 |  |
| 1st base |  |  |  | C | 82 |  |
| 2nd base |  |  |  | D | 72 |  |
| Shortstop |  |  |  | E | 13 |  |
| 3rd base |  |  |  | F | 77 |  |
| Left field |  |  |  | G | 24 |  |
| Center field |  |  |  | H | 80 |  |
| Right field |  |  |  | I | 85 |  |
| DH |  |  |  | J | 65 |  |
| PH |  |  |  | K | 81 |  |
| **Total** |  |  |  | **Total** | 721 |  |

Your task is to determine from the offensive data alone which position is which. Work in groups of 3 or 4 to develop a model of how offensive production should depend on field position. Why should it depend on field position at all? Feel free to use any plotting techniques that might be helpful in developing the model. Then compare your model to the actual data. Does the model need to be revised? Can you determine which position is which?

**Two-Way Frequency Table**

A two-way frequency table lists data that can be **summed in two directions**. A two-way table is used to study the relationship between two categorical variables. As with a one-way frequency table, entries can be given as frequencies, relative frequencies, or percentages.

**Example:** Preferred color of 20 1st graders by gender (redux-redux)

Which color do you like better, red or blue?

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Red** | **Blue** | **Gender Total** |
| **Girls** | 5 | 7 | 12 |
| **Boys** | 2 | 6 | 8 |
| **Color Total** | 7 | 13 | 20 |

**Marginal**

**Frequencies**

**Joint**

**Frequencies**

What is the relative frequency in the class of girls who prefer red?

What percentage of the class are boys who prefer blue?

**Two-way frequency tables are useful for determining conditional probabilities.**

Probability is the *expected percentage* for a given outcome.

What is the probability that a boy will like red?

What is the probability that a 1st grader who likes blue is a girl?